Studies in History and Philosophy of Science 117 (2026) 102151

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives

Arno Simons *, Michael Zichert, Adrian Wüthrich

History and Philosophy of Modern Science, Technische Universität Berlin, Straße des 17. Juni 135, 10623, Berlin, Germany

ARTICLE INFO

Keywords:
History philosophy and sociology of science (HPSS)
Large language models (LLMs)
Digital humanities
Conceptual analysis
Historiography
Qualitative research

ABSTRACT

This paper examines large language models (LLMs) as research tools in the history, philosophy, and sociology of science (HPSS). Because LLMs can work directly with heterogeneous, unstructured texts and capture meaning-relevant associations from usage patterns, they offer new ways to bridge close reading and corpus-scale analysis, challenging the idea that computational scale and interpretive nuance must trade off. We provide a compact primer on LLMs, covering the main components of their neural network architecture, the differences between generative and full-context models, and adaptation strategies such as fine-tuning, prompt-based learning, and retrieval-augmented generation (RAG). Building on this foundation, we analyze how LLMs recast three classic methodological problems in HPSS: working with historically messy data, detecting and interpreting large-scale patterns, and modeling scientific change over time. Across these areas we synthesize recent work in HPSS and adjacent fields, and we clarify how LLM outputs can function as exploratory prompts, as inputs to more structured pipelines, or as evidence under stricter validation and documentation. We conclude with four lessons: 1) model choice embeds interpretive trade-offs, 2) responsible use requires LLM literacy, 3) HPSS should develop its own tasks and evaluation practices, and 4) LLMs should extend rather than replace established interpretive methods. We also situate these methodological questions within broader concerns about platform dependence, accountability, and the responsibilities attached to research infrastructures. Finally, we argue that HPSS is well positioned to both use LLMs and to interrogate what counts as explanation, evidence, and responsible use in interpretive research.

When citing this paper, please use the full journal title *Studies in History and Philosophy of Science*.

1. Introduction

History, philosophy, and sociology of science (HPSS), which focuses on the historical development, conceptual foundations, and social organization of science, has long treated knowledge as context-dependent and theory-laden. As in the humanities more generally, computational methods in HPSS have often been met with skepticism, seen as trading nuance for scale or sacrificing contextual richness for abstraction (e.g., Buchholz & Grote, 2023; Da, 2019). When *Isis* launched its 2019 Focus section on Computational History and Philosophy of Science (Gibson et al., 2019; Laubichler et al., 2019), contributors pointed to both promise and persistent obstacles: challenges in curating structured data, the need for domain-sensitive tools, and difficulties linking statistical patterns to historical meaning. Although the "computational turn" aimed to bridge close reading and large-scale analysis, the divide remains, sustained by technical barriers and epistemic concerns (e.g., Leydesdorff et al., 2020). Yet amid these tensions, authors in this journal have called for more integrative methodological frameworks that combine computational tools with philosophical inquiry (Herfeld & Lisciandra, 2019) and for bridging qualitative and quantitative methods more generally (Hangel & ChoGlueck, 2023), a direction that LLMs may now help to advance.

LLM may mark an inflection point in the tension between computational and interpretive approaches, improving on earlier computational methods in three ways. First, they are more accessible, working with relatively unprocessed, heterogeneous texts and following ordinary-language instructions with less preprocessing or coding. Second, they are more versatile because a single system can support many

* Corresponding author.
E-mail address: arno.simons@gmail.com (A. Simons).

https://doi.org/10.1016/j.shpsa.2026.102151
Received 14 June 2025; Received in revised form 28 January 2026; Accepted 15 March 2026
Available online 30 March 2026

activities that previously required separate tools and workflows. Beyond generating fluent prose for tasks such as thematic exploration, LLMs can also produce structured outputs like category or entity labels and similarity scores that enable comparisons between texts or terms. Third, they support meaning-sensitive analysis more directly by better capturing contextual nuance and polysemy, and by producing outputs, textual and numerical, closer to interpretive practice.

Because LLM outputs can resemble familiar qualitative products like annotations, codes, memos, or hypotheses, they also raise methodological and epistemic questions. Some echo long-standing qualitative concerns, including transparency, reflexivity, and the lack of clear "ground truths". Others are specific to LLM workflows and center on warrant and trust: is an LLM-generated output evidence, interpretation, a testable hypothesis, or a cue for further reading? Since outputs can vary with phrasing and often sound more certain than sources allow, they should be documented, stress-tested with prompt changes, and treated as provisional unless confirmed through close reading or other evidence.

Against this backdrop, LLMs expand the HPSS toolkit by easing movement between qualitative and quantitative work and by supporting analysis of heterogeneous sources at multiple scales. For historians of science, they can support more systematic tracing of conceptual change by using similarity measures to detect shifts in how target terms are deployed across periods, genres, and institutions, while keeping relevant passages close for verification. For philosophers of science, they can support the analysis of concepts and arguments, generate rival positions to surface implicit assumptions, and clarify commitments. For sociologists of science, they can help identify and compare passages related to authority claims, boundary-work, or controversies, both through structured classification and open-ended responses.

In this paper, we examine the opportunities and challenges LLMs pose for HPSS research, focusing on their use as research tools rather than as writing aids or general-purpose assistants. We aim to bridge computational and interpretive perspectives by pairing accessible explanations of LLM methods with critical reflection on their epistemic and methodological implications. Section 2 provides a short primer on LLMs, covering contextualized word embeddings, key architectural differences, the accessibility–literacy trade-off, and LLMs as an operationalization of distributional semantics. Sections 3–5 then examine how LLMs recast three classic methodological problems in HPSS: working with historically messy data (Section 3), detecting and interpreting large-scale patterns (Section 4), and modeling scientific change over time (Section 5). Section 6 distills four lessons for responsible integration, and Section 7 concludes.

2. A short primer on LLMs

By LLMs we mean neural language models based on the transformer neural network architecture (Vaswani et al., 2017). Besides generative models, such as GPT (Radford et al., 2018), this also includes full-context models, such as BERT (Devlin et al., 2018). In this section we will explain how token embeddings underpin how both model types represent and process text, how differences in their architecture and training influence behavior, and how these choices influence both the capabilities and limitations of LLMs when working with text. Readers already familiar with transformer models may wish to skip to the HPSS-relevant applications in the next sections.

2.1. Embeddings and architecture

Inside an LLM, text is broken into tokens, which may be whole words, subword pieces, or symbols. Each token is mapped to an initial, static input embedding, with its position encoded too. These vectors are then updated through a stack of transformer layers. In each layer, "self-attention" lets every token incorporate information from other tokens in the input by weighting them for relevance to the current token's update. The model learns the parameters that encode these relevance patterns during training, while the attention weights are computed anew for each input. While "attention" is a conventional name for this mechanism, the metaphor is useful only insofar as it conveys selective routing of information among tokens in the context, but it breaks down if taken psychologically. The mechanism is not conscious, goal-directed, or voluntarily steerable.

Across layers, each token's representation is repeatedly updated, yielding one new context-sensitive vector per token per layer. "Context-sensitive" here means that the same word or subword unit can receive different representations depending on the surrounding text in a given input (in contrast to static input embeddings), while "across layers" means that within a single input the representation of a token is updated from one transformer layer to the next. These layer-wise, context-sensitive token representations are often referred to as contextualized word embeddings (CWEs). Empirically, different kinds of linguistic information tend to be more recoverable at different depths, with earlier layers often reflecting more local or syntactic cues and later layers reflecting more aggregated semantic and discourse-level cues (e.g., Ethayarajh, 2019; Jawahar et al., 2019).

Both static input embeddings and layer-wise CWEs are learned under particular training objectives on particular data. The patterns they encode reflect what was useful for that training setup, not necessarily the semantic distinctions an HPSS analysis cares about. Whether they fit a given interpretive target is therefore an empirical question.

2.2. Full-context vs. generative models

While LLMs are often associated with interactive dialogue, especially through chatbots like ChatGPT, not all are designed for text generation. Broadly, LLMs can be grouped into generative and full-context models. The distinction originates in their pretraining objectives and attention patterns, but behavior in practice also reflects later stages such as fine-tuning and instruction- or preference-tuning, which influence how models are used downstream. Table 1 summarizes the key architectural and functional differences between these two model types.

Generative models such as GPT (Radford et al., 2018) are trained "autoregressively" to predict each next token from the preceding context, so CWEs are learned within a left-to-right sequence. This objective makes them especially effective at producing fluent, coherent text. As these models are scaled in parameters, data, and compute and then instruction- and preference-tuned, they show strong zero- and few-shot performance, where prompts and occasional examples specify the task, and improved multi-step performance often framed as "reasoning", especially when prompted to write intermediate steps (Wei

Table 1
Key differences between full-context and generative LLMs.

	Full-Context LLMs (e.g., BERT, SciBERT)	Generative LLMs (e.g., GPT-54, Claude-4.5)
Architecture & Pretraining	Bidirectional (masked token prediction)	Autoregressive (left-to-right token prediction)
Primary Output Type	Token and text embeddings, classification scores, token-level predictions	Coherent text generation, structured outputs via natural language prompting
Representative HPSS Applications	Conceptual history, scientific entity extraction, citation content classification, research topic modeling	Few-shot learning of HPSS related tasks, prompting for hypothesis generation, interactive retrieval with source-linked excerpts (RAG), synthetic data generation
Accessibility	Moderate to low: requires technical setup, local compute, and literacy	High: accessible via web interface or API, minimal setup required
Transparency & Platform Dependence	More frequently open-source, models and training data often inspectable and reusable	Often proprietary and opaque, limited insight into training data or internal parameters

et al., 2022). Many deployed systems additionally support "tool use", letting models call external resources like search, calculators, or databases to access up-to-date information or exact results.

Full-context models such as BERT (Devlin et al., 2018) use masked language modeling, replacing random tokens with "[MASK]" and predicting them from bidirectional context. Because attention spans the whole input, each token's CWE is determined by both left and right context. In practice, these models are typically fine-tuned for downstream tasks via small task-specific output layers (e.g., entity recognition or sequence classification), or adapted with contrastive objectives to produce fixed-length text embeddings for similarity search and retrieval (see Section 3.4). This makes them well suited for structured prediction and representation-focused applications, and they perform strongly on meaning-sensitive benchmarks (e.g., Clark et al., 2020; Devlin et al., 2018; Liu et al., 2019; Reimers & Gurevych, 2019).

Since 2018, LLMs have diverged in scale and use. Generative models like GPT-5 have grown to trillions of parameters and dominate commercial applications due to their fluency and generalization. However, they are typically closed-source and resource-intensive. Full-context models, by contrast, are smaller, often open-source, and accessible for local deployment. While less flexible, they are sometimes more useful in structured, non-generative tasks.

2.3. The accessibility–literacy trade-off

Another key difference between generative and full-context LLMs concerns how easily researchers can put them to work. Generative models are typically packaged for ease of use. They support "in-context learning", where models perform tasks by receiving instructions (zero-shot) or examples (few-shot) directly in the prompt, without the need for fine-tuning. Many also bundle multimodal inputs and tool use like code execution, APIs, or web search. But this convenience can also compress consequential choices into largely hidden defaults (post-processing, prompt templates, retrieval and ranking), and fluent outputs can sound more warranted than the cited or retrieved evidence supports.

Full-context models more often appear as components in task-specific pipelines. Using them typically involves technical setup, local compute, and choices about fine-tuning, post-processing, and evaluation. This raises the barrier to entry, but once running, these workflows can make intermediate artifacts more inspectable, for example embeddings, neighbors, clusters, retrieval results, and decision thresholds, depending on the pipeline.

In practice, this yields an accessibility–literacy trade-off. Generative models lower the barrier to entry, but do not lower the literacy required for responsible interpretation. Full-context models raise the barrier to entry (setup, compute, pipelines), but can increase inspectability once running.

2.4. LLMs as an operationalization of the distributional hypothesis?

LLMs operationalize a distributional approach to language (see Brunila & LaViolette, 2022; Grindrod, 2023). Trained primarily on text to predict tokens from preceding (generative models) or surrounding (full-context models) context, they represent linguistic units in terms of how those units pattern across linguistic environments. This echoes Harris' (1954, p. 156) claim that "differences in meaning correlate with differences in distribution".[1] The strong performance of both generative and full-context LLMs on many tasks that probe meaning-relevant distinctions (see Section 2.1) extends, and provides new empirical support for, earlier results in NLP and computational linguistics (e.g., Landauer & Dumais, 1997; Mikolov, Sutskever, & Chen, 2024; Turney & Pantel, 2010) showing that distributional patterns carry substantial information about many distinctions language users treat as semantic. In Grindrod's (2023) terms, such results are straightforwardly supportive of the weaker distributional hypothesis: a systematic correlation between meaning and distribution. The stronger thesis, which holds that distributional properties place a constitutive constraint on any adequate theory of meaning, though not necessarily that they exhaust meaning, is a further and contested step.

While none of this implies that LLMs capture more than distributional aspects of meaning (cf. Grindrod, 2023), and benchmark success does not by itself identify what meaning-relevant property a model is tracking, there is ongoing debate about whether capacities often treated as "non-distributional" relative to text-only co-occurrence, including some pragmatic inferences and grounding-by-proxy, can nevertheless be learned from text (e.g., Bender et al., 2021; Bubeck et al., 2023). Relatedly, discussions of a broader notion of distributional learning over extra-linguistic variables, which could in principle subsume further aspects of grounding and pragmatics (Brunila & LaViolette, 2022; Grindrod, 2023), have been linked to the prospect of multimodal LLMs that learn joint regularities beyond text-only co-occurrence. Regardless of where one lands in these debates, the demonstrated capabilities of current models suggest that they recover enough meaning-relevant structure to make a systematic discussion of opportunities, limits, and risks for HPSS both timely and important (cf. Simons, Wüthrich, et al., 2026). Throughout, we therefore focus on how these capacities can be put to work for HPSS aims and where they may fail, and we avoid anthropomorphic formulations that imply human-like understanding or agency.

3. Data, models, and training

In LLM-based research workflows, the role of data shifts in subtle but significant ways. Rather than requiring fully structured inputs from the outset, these models are designed to learn and process both syntactic and semantic regularities from unstructured text. As a result, interpretive choices increasingly take place within model design, training, and prompting, reframing traditional challenges of data curation as questions of model adaptation and interpretive fit.

3.1. Data complexity and model assumptions

Advocating the computational turn in HPSS just before the rise of LLMs, Laubichler et al. (2019) identified the provision and curation of structured data as a central challenge. HPSS data, they argued, are often fragmented, inconsistently formatted, and hard to structure without sacrificing nuance. More than a technical issue, this is an epistemological one: knowledge is historically situated, and so are the categories, concepts, and evidentiary standards embedded in data. What counts as meaningful shifts over time, conditioned by changing scientific practices and cultural contexts. Data curation must therefore contend with both the fragmented form and historical content of knowledge.

LLMs appear to shift some of these challenges by working well with unstructured text, processing raw language without predefined categories or formats. LLMs can also be used to generate synthetic data, for example to alleviate data scarcity in historically specific corpora (Danilova et al., 2026). But this flexibility brings new complexities. The model itself becomes an epistemic infrastructure: not just a tool trained on data, but a condensed representation of large text corpora, formed through decisions about inclusion and encoding, and embedded in the workflows, assumptions, and institutions that structure knowledge production. This raises concerns for HPSS, where much of the material comes from earlier periods. LLMs trained mainly on contemporary data may flatten or misrepresent historically specific language and concepts, particularly in archival texts, outdated vocabularies, or shifting

[1] Alongside Harris, Firth's slogan that "the meaning of a word is the company it keeps" is often treated as a paradigmatic formulation of the distributional hypothesis. For discussion of key differences between Firth's and Harris' accounts, and of how these differences matter for NLP and contemporary LLM-based approaches, see Brunila and LaViolette (2022).

conceptual frameworks (cf. Simons, Wüthrich, et al., 2026, part 2 "Historicizing LLMs").

As a result, we must attend not only to data, but also to how models are trained and what assumptions they encode. At the same time, structured data remains crucial, not just for fine-tuning or grounding LLMs via sources like knowledge graphs or citation networks, but also as a primary object of inquiry. HPSS scholars will continue to build and analyze structured datasets or knowledge graphs, e.g. to trace disciplinary development, map intellectual networks, and make historical claims beyond what LLMs can currently infer. While LLMs can assist with scaling curation and extraction of such data from unstructured text (Boulanger, 2026; Schlattmann et al., 2026), the interpretive work of deciding what to represent, how to operationalize categories, and how to analyze the resulting structures remains central. The data challenge has not disappeared. It has changed form.

HPSS researchers work with diverse materials, from correspondence, policy texts, and media discourse to field notes, interviews, and multimodal artifacts (figures, diagrams, images, audio). These sources bring distinct complexities, including temporal shifts in language and conceptual frameworks. Because LLMs operate mainly on text, their usefulness depends on how well the available inscriptions fit the phenomenon under study. When the phenomenon is itself inscriptional, LLMs can directly support analysis, though interpretive assumptions about context and meaning still matter. When inscriptions are traces of broader activity, LLMs can help organize and analyze them, but stronger assumptions are needed to infer underlying practices. Multimodal models may broaden what can be analyzed (see Section 2.4), but remain immature (e.g., Wu et al., 2023). Across settings, we recommend triangulating LLM outputs with established qualitative methods to strengthen validity, while recognizing that LLM-based analysis can still be informative on its own.

3.2. Domain-specific pretraining

A key strategy for adapting LLMs to specialized domains is domain-specific pretraining, where models are exposed to targeted corpora during their initial learning phase. This structures internal representations and constrains the embedding space in ways that persist, influencing how models process and generate language. While full pretraining is resource-intensive, a common alternative is continued pretraining: further training an existing model like BERT on domain-specific texts that were underrepresented in the original corpus. Pioneered by models like BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019), this approach has also been explored in a small number of HPSS case studies (Simons, 2024a; Zichert et al., 2025), which suggest it can be useful in some settings, though the evidence base is still limited and likely to be context-dependent.

Pretraining a model from scratch on scientific data was first explored with a variant of SciBERT and later adopted by models like PubMedBERT (Gu et al., 2021), BioGPT (Luo, Sun, et al., 2022), and Galactica (Taylor et al., 2022). This approach offers greater adaptability by avoiding biases from general-purpose models. Unlike continued pretraining, which retains the base model's vocabulary, from-scratch pretraining enables vocabulary customization to better represent specialized terms. It also allows architectural changes, such as adapting attention mechanisms for temporal information (Rosin & Radinsky, 2022; cf. Büttner, 2026). However, given its high data and computational demands, from-scratch pretraining remains impractical for most HPSS applications. For a comprehensive overview of targeted pretraining for scientific texts, see Ho et al. (2024) and Zhang et al. (2024).

3.3. Task-specific fine-tuning

Beyond pretraining, LLMs can be adapted to HPSS-specific purposes through fine-tuning on particular NLP tasks. In NLP, a "task" refers to a defined goal, such as sentence classification, question answering, or named entity recognition, where the model must act on text in specific ways. Fine-tuning for such tasks typically involves supervised learning on datasets that pair inputs with expected outputs. This requires carefully labeled examples, often informed by human labor and interpretive judgment. In HPSS contexts, where interpretations are historically situated and categories are fluid, the assumptions embedded in labeled data carry particular weight.

Three widely used strategies include adding task-specific classification layers, applying prompt-based fine-tuning, and using contrastive learning, as exemplified by science-specialized models such as BioBERT, BioGPT, and SPECTER (Cohan et al., 2020), respectively. Table 2 outlines how these approaches fit into the wider training landscape and highlights potential HPSS applications.

Task-specific classification typically adds one or more layers to a pretrained model to map token or span representations to predefined labels. BioBERT uses this approach for biomedical tasks such as named entity recognition and relation extraction, learning to identify patterns like gene–disease associations in annotated data. These added layers transform CWEs into task-specific predictions, guided by labeled training examples.

Prompt-based fine-tuning treats structured tasks as text generation problems, using natural language prompts instead of added output layers (Liu et al., 2023). The task is encoded directly in the input, typically by researchers using templates. For example, BioGPT was trained on pairs like: Input: "What is the relationship between aspirin and COX-1?"; Output: "Aspirin inhibits COX-1".

Contrastive learning fine-tunes full-context models on text pairs labeled as similar or dissimilar to produce fixed-length text embeddings that support tasks like document clustering or retrieval. SPECTER, for example, builds on SciBERT and was trained using citation links as a proxy for similarity, avoiding manual labeling, but inheriting whatever kinds of similarity citations actually encode (see the next section). Text embeddings from models like SPECTER or more advanced Sentence-Transformers (Reimers & Gurevych, 2019) now underpin many methods entering HPSS, including thematic clustering and novelty detection.

By relying on labeled training data, all strategies mentioned in this section encode specific assumptions about semantic similarity and relevance, which carry particular weight in HPSS contexts. When applying tools like BioBERT, BioGPT, or SPECTER to historical inquiry,

Table 2
Strategies for adapting LLMa to HPSS research contexts.

	Core Idea	Data Required	Representative HPSS Application
Domain-Specific Pretraining	Exposing models to HPSS-specific language during pretraining	Large domain-specific corpus	Capturing field- or time- specific semantics in scientific texts
Task-Specific Fine-Tuning	Training a model on supervised data for classification, NER, etc.	Labeled examples per task	Scientific entity and relation extraction, citation context classification, argument mining
Contrastive Fine-Tuning	Optimizing pooled embeddings by training on similar/dissimilar sequences	Similar/dissimilar sentence pairs or documents	Research topic modeling, novelty detection, revision tracking
Prompt-Based Learning	Framing tasks as instructions or examples in prompts	None or a few illustrative examples	Few-shot learning of HPSS related tasks, prompting for hypothesis generation, synthetic data generation
Retrieval-Augmented Generation (RAG)	Augmenting generation with external document retrieval	External corpora + retrieval embeddings	Interactive sparring with source-linked excerpts (RAG)

researchers must account for the temporal mismatch between the models' training data and their sources. And as discussed next, common similarity metrics driving text embeddings may diverge from the interpretive concerns of a given HPSS investigation.

3.4. Text embeddings: design choices and interpretive consequences

In contrast to CWEs, which learn token-in-context representations from pretraining on naturally occurring text (though still constrained by corpus selection and objectives; see Section 2.1), text embeddings learn a task-specific notion of similarity by training models to place whole texts close or far apart using constructed positive and negative pairs, drawn from manual labels or from proxies in "self-supervised" setups, such as citations or hyperlinks.

For HPSS, this means that text embedding spaces can hard-code a proxy for relatedness that may diverge from the analytic target. For example, novelty measures based on embedding distance (Section 5.2) using models trained on citation signals (e.g., SPECTER) may reflect not only conceptual difference, as is their target, but also other things that citation practices encode, including norms, ties, and rhetoric. More broadly, embedding-based analyses always operationalize "relatedness" in specific ways and thus carry interpretive assumptions that make some relations visible and others harder to see.

3.5. Retrieval-augmented generation (RAG) and tool use

Beyond pretraining and fine-tuning, retrieval-augmented generation (RAG) and other "tool-based" approaches offer flexible ways to adapt LLMs to HPSS research. In RAG, a generative model is combined with an external retrieval system, often using text-embedding-based similarity, to fetch relevant texts that are then used in the model's output (Lewis et al., 2020). Tool-based systems more generally let models call external resources, such as search or databases, which can improve access to sources and traceability. As these systems spread in domain tools and platforms like ChatGPT, Gemini, and Claude, they support natural-language interaction with sources that goes beyond keyword search. In HPSS they can enable iterative exploration or "interpretive sparring" (Hill, 2026; Scharnhorst et al., 2026; Tykhonov et al., 2025).

But these affordances also introduce epistemic risks. With more functionality embedded in natural language interfaces, interpretive assumptions are often obscured, embedded within retrieval algorithms, similarity metrics, or prompt templates. This opacity may conflict with HPSS values, reinforcing the need for transparency and critical scrutiny in how such systems are adopted and used.

Overall, Laubichler et al.'s (2019) data challenges have shifted rather than disappeared. LLMs ease structuring demands but sharpen concerns about transparency, reproducibility, and infrastructural control: proprietary models limit scrutiny, while open-source options require unevenly distributed compute. This makes LLM literacy and governance central, and foregrounds questions of accountability and research infrastructure politics.

4. Patterns

This section addresses Laubichler et al.'s (2019) second key methodological challenge: how to detect and interpret patterns in scientific knowledge at scales beyond traditional close reading. LLMs extend earlier computational methods by surfacing corpus-scale structure while retaining more local context, including taxonomies, disciplinary classifications, themes, and genre features such as headings, citation practices, metadata, and authorship conventions. They support tasks like topic modeling, named entity recognition, and relation extraction, but outputs reflect model representations and training data, and generative models add risks of prompt sensitivity and hallucination. Recent work applies these methods across scales, from words to documents, in two broad modes: exploratory pattern discovery and targeted detection of predefined categories.

4.1. Exploration of unknown patterns

As noted above, CWEs capture statistical patterns of usage that often track both syntactic functions and context-sensitive semantic distinctions (see Section 2.1). This makes them particularly valuable for HPSS research concerned with conceptual variation and the situated use of language. At the token level, CWEs can be used to quantify how stable or variable a term's usage is across contexts.

Despite their potential, such methods remain underused in HPSS contexts. Kleymann et al. (2022) applied CWE clustering to the word "theory" across nearly 4000 digital humanities articles. Although the study aimed to identify sense clusters, the resulting groupings primarily reflected syntactic variation rather than the semantic distinctions the authors were looking for. Simons (2024b) evaluates general vs domain-adapted BERT models on about 4000 manually labeled occurrences of the single target term "Planck", drawn from two corpora: a 1500-paragraph Astro-HEP arXiv sample (2900 occurrences) and a physics-Wikipedia corpus built from 6642 articles (1186 occurrences across 885 paragraphs). In this single-term case study, the clusters aligned reasonably with anticipated sense distinctions, and the domain-adapted models appeared to separate some distinctions more clearly. How reliably this carries over to other target terms, periods, or corpora remains an open empirical question. Both Kleymann et al. (2022) and Simons (2024b) also connect to token-level tracing of change over time, which we return to in Section 5.1.

To study conceptual variation in scholarly jargon, Lucy et al. (2023) used a substitute-based word sense induction method. ScholarBERT generated the top five substitutes for 4000 lemmatized target words, which were used to build co-occurrence graphs and clustered into word-sense candidates via Louvain community detection. They recovered clusters interpretable as discipline-specific senses for many terms, such as "bias" in statistics, psychology, and climate science, but noted challenges with clustering granularity and high computational demands.

A closely related strategy focuses less on producing discrete senses and more on measuring contextual stability as a signal of how standardized a term's usage is within a discourse community. Ahmadi (2026) does this with a Semantic Uniformity Score (SUS) derived from BERT contextualized word embeddings. She computes within-term similarity across all occurrences of each token that passes her frequency and preprocessing filters, then aggregates these scores to compare astrophysics and sociology, finding higher semantic uniformity in the former, which she interprets as stronger linguistic codification. She also cross-checks the dispersion results by clustering embeddings for a small set of terms ("order", "paradigm", "Planck", and "wave") and inspecting the resulting groupings.

For historians of science, these techniques could, in principle, help identify divergent uses of key terms across corpora, and they can support mapping conceptual variation across disciplines or schools of thought. Philosophers may use them to analyze shifts or ambiguities in the usage of foundational concepts within theoretical debates, whether by inspecting candidate sense groupings or by tracking when a term's usage becomes more or less contextually stable (cf. Malaterre & Lareau, 2026). For sociologists, these methods may offer a way to detect terminological distinctions that may correlate with institutional affiliations or disciplinary boundaries, and to operationalize degrees of linguistic codification while remaining attentive to the interpretive limits of any single metric or clustering outcome.

Moving beyond token-level analysis, a core strategy for exploring semantic patterns in unstructured text using LLMs involves clustering fixed-length text embeddings to identify latent topical structure. This approach uses sentence- or document-level embeddings generated by models such as Sentence-BERT or SPECTER and applies clustering algorithms to uncover emergent groupings based on semantic similarity.

Tools like BERTopic (Grootendorst, 2022), which combine these embeddings with dimensionality reduction and topic ranking, have become especially popular due to their flexibility and accessible design. BERTopic has emerged as a serious alternative to existing topic modeling technologies, offering comparable or superior topic coherence in many settings. In HPSS, it has been applied to map thematic structures in scientific corpora (e.g., Kim et al., 2024) and public discourse (e.g., Falkenberg et al., 2022), as well as to assess temporal representations in synthetic historical data (Danilova et al., 2026). We return to further HPSS-specific applications in Section 5.2, where we discuss BERTopic's use in modeling discursive change over time.

While BERTopic currently dominates LLM-based topic modeling in HPSS contexts, several alternatives have been explored. These include hybrid models that combine BERT embeddings with traditional approaches like LDA or variational autoencoders (Bianchi et al., 2021; George & Sumathy, 2023), as well as FASTopic (Wu et al., 2024), which links documents, topics, and words semantically. BERTopic performs comparably to, or better than, VAE-based models in terms of topic coherence, while being simpler and more efficient (Grootendorst, 2022; Zhang et al., 2022). Yet its assumption of a single dominant topic per document can obscure conceptual multiplicity (Egger & Yu, 2022), and its results are sensitive to clustering parameters and dimensionality reduction. As with all embedding-based methods, it also inherits biases from general-purpose training corpora, which may misrepresent historical language or domain-specific concepts. FASTopic promises richer topic distributions and improved interpretability, while newer generative models like TopicGPT (Pham et al., 2023) and PromptTopic (Wang, Prakash, et al., 2023) allow more flexible, human-readable output, though they come with higher computational cost. Looking ahead, hybrid frameworks that combine embeddings, prompting, and probabilistic modeling may hold promise for HPSS applications.

4.2. Detection of known patterns

In contrast to exploratory approaches, which let patterns emerge from data, many HPSS-relevant tasks involve the detection of predefined categories or structures. These range from token-level classification (e.g., word sense disambiguation), to span-level detection (e.g., citation context extraction), to document-level classification (e.g., rhetorical function or disciplinary field). Combined appropriately, these techniques can recover complex structures such as entity networks, argumentative sequences, or causal relations. To perform these tasks, researchers typically fine-tune pretrained LLMs on labeled datasets, especially full-context models like SciBERT. Alternatively, generative models such as GPT-5 can be prompted directly in few-shot or zero-shot settings, enabling task performance without extensive retraining.

Citation context analysis (CCA) is a prominent use case, especially within bibliometrics, since it maps cleanly onto span detection and context classification tasks that can be handled with either fine-tuned encoder models or prompted generative models. To classify citation context using one or more labels to indicate what a citation is doing, for example rhetorical function, evaluative stance, hedging, or perceived importance, scholars have used fine-tuned full-context classifiers (e.g., Beltagy et al., 2019; Nicholson et al., 2021), unsupervised pipelines that cluster text embeddings of citation contexts and then map clusters to intent labels (Roman et al., 2021), as well as generative models (e.g., Arnaout et al., 2025; Kunnath et al., 2023, pp. 1127–1137). To delineate the citation context before classification, moving beyond rule-of-thumb windows such as “use the citing sentence”, researchers have proposed both supervised sequence-to-sequence attention models that generate a minimal reference-text fragment from the citation sentence (Khan et al., 2025) and unsupervised dynamic context extraction pipelines that expand the context by selecting neighboring (optionally non-contiguous) sentences whose embeddings are most similar to an embedding of the cited paper, before passing the resulting span to a downstream classifier (Kunnath et al., 2022).

For HPSS, model choice is only part of the story: the assumptions carried by training data and annotation guidelines, prompt framing, context delineation, and label design strongly shape what the outputs can legitimately support (cf. Simons, Arnaout, & Gurevych, 2026). This matters because fine-tuned models reproduce the interpretive constraints and defaults of the datasets they learn from, while prompted generative models make those constraints more fluid but also more dependent on the prompt and harder to standardize across runs. As Liesegang & Gläser (2026) argue, LLM support for CCA is therefore a double-edged sword: it can scale detection and classification of citation contexts and force implicit decisions to be made explicit, but it can also harden fragile assumptions about context, meaning, and exhaustiveness into systematic error unless those assumptions are critically examined and validated.

Beyond citation analysis, LLMs have been widely used in scientific domains to extract information like domain-specific entities or material–property associations (Dagdelen et al., 2024; Ji et al., 2020). They have also been applied to map argumentative and causal dependencies (Fergadis et al., 2021, pp. 100–111; Gorur et al., 2024; Zhang et al., 2024), and to classify documents by topic, contribution type, or structural role (Chen et al., 2022a, 2022b; Ma et al., 2022). Across these applications, fine-tuned domain-specific models like BioBERT, SciBERT, and MatBERT consistently outperform general-purpose baselines on such tasks (Beltagy et al., 2019; Dagdelen et al., 2024; Ji et al., 2020; Lee et al., 2020; Luo, Sun, et al., 2022). GPT-style models, in contrast, perform well in zero- and few-shot scenarios when given clear prompts. Their advantages lie in accessibility and adaptability; their weaknesses include domain specificity, factual unreliability, contextual ambiguity, high computational costs, and integration complexity (Shao et al., 2024; Zhu et al., 2024). Since scientific reasoning in many fields is inherently multimodal, LLMs and pipelines that integrate textual and visual signals may improve robustness for scientific information extraction. Current models show promising capabilities in simple cases but are still unreliable in complex tasks (Alampara et al., 2025).

HPSS scholars have recently begun to apply these methods, or combinations of them, to extract structured data from unstructured HPSS corpora, addressing one of the core challenges highlighted by Laubichler et al. (2019). Boulanger (2026) uses LLM-assisted extraction to identify citation and bibliographic entities in socio-legal and humanities materials as a step toward building a disciplinary-history knowledge graph. Schlattmann et al. (2026) combine LLM-based entity and relation extraction with human review to turn biographical lexicons into an ontology-guided historical knowledge graph. Together, they highlight opportunities to open under-indexed corpora and support new kinds of cross-text comparison, while also underscoring familiar challenges around messy metadata, interpretive ambiguity, and the need for reliable validation and shared standards.

The LLM-based techniques for pattern detection discussed in this section offer powerful tools for surfacing conceptual, rhetorical, and thematic structures across scientific texts. In structured tasks with limited ambiguity, LLMs may support relatively robust, automatable insights. In more open-ended interpretive contexts, they can surface plausible framings or discursive patterns that warrant further analysis. Their value lies not in offering definitive classifications, but in expanding the range and scale of what can be noticed, compared, and questioned.

5. Dynamics

The third major methodological challenge identified by Laubichler et al. (2019) is explaining scientific change. While the earlier challenges—data structuring and pattern detection—concern how knowledge is represented and recognized, this one asks how it evolves: how concepts, practices, and institutions emerge, shift, or dissolve over time. For HPSS scholars, such questions have traditionally been addressed through contextual analysis and critical source work. This section asks whether

and how LLMs might contribute to that task.

We distinguish two complementary levels at which LLMs can model scientific change: token-level dynamics, capturing shifts in the use and semantic associations of individual terms, and text-level dynamics, modeling broader discursive patterns such as topic emergence, argumentative change, or evolving citation practices. At both levels, LLM-based methods offer new ways to access diachronic patterns, surfacing trends that might escape close reading. But as with other challenges identified by Laubichler et al., these affordances come with epistemic costs: issues of interpretability, operationalization, and historical alignment persist. What counts as semantic change? When does a textual shift reflect a conceptual rupture? And can statistical similarity serve as a proxy for historical continuity?

The sections that follow survey LLM-based approaches to both levels, examine their assumptions, and assess their relevance for HPSS research.

5.1. Token-level dynamics

At the most granular level, token-level analysis can trace how individual lexical items shift in usage and in their distributional semantic associations over time. By comparing CWEs across temporally segmented corpora, researchers can apply lexical semantic change detection (LSCD) (Periti & Montanelli, 2024) to reconstruct conceptual histories (Zichert & Simons, 2026) and to quantify shifts in disciplinary vocabularies. These models build on earlier co-occurrence and distributional methods (Gavin et al., 2016; Kutuzov et al., 2018; Wevers & Koolen, 2020), while offering finer-grained, context-sensitive representations of distributional meaning.

Recent HPSS applications reveal both the potential and limits of these methods. Two studies already introduced in Section 4.1 for word sense modelling also illustrate how CWE-based analyses can support term-level tracing. Kleymann et al. (2022) track the evolution of the concept of *theory* in digital humanities writing by fine-tuning a BERT model on the journal corpus, extracting CWEs for occurrences of "theory" and related epistemic terms such as "model" and "method", and comparing period-specific and aggregated representations via cosine similarity. This makes it possible to trace shifts in a term's semantic neighbourhood, and with it changes in use that invite interpretation as conceptual change. Simons (2024b) likewise shows how CWE-based analyses can support focused diachronic tracing when distributional shifts are triangulated with historical knowledge and qualitative inspection, even in single-term case studies. Complementing these term-focused approaches, Zichert et al. (2025) trace the conceptual history of the *virtual particle* by using "virtual" as a linguistic marker and analysing its usage across a large physics corpus. They track shifts in dominant usage patterns as well as changes in polysemy, and complement this approach with dependency parsing, which identifies the nouns most often used with "virtual" and serves as an interpretive cross-check on the LLM-based methods.

While all of the HPSS case studies make use of full-context models, generative models have recently entered LSCD research along two lines. First, they are used directly for change detection through prompting, for example via sense judgements or lexical substitute tasks. Periti et al. (2024) show that while such approaches are promising, encoder-based models still perform better, especially for fine-grained, short-term distinctions. Second, generative models are increasingly used in hybrid workflows, where they generate historically plausible, sense-specific examples or definitions that serve as synthetic diachronic data, while change is still measured with encoder-based models (Cassotti & Tahmasebi, 2025a; 2025b). Together, all these case studies show how LLMs can support historically grounded concept tracing, but that such approaches need to be paired with careful corpus, dataset, and model selection, as well as evaluation practices that combine close reading and expert validation with established quantitative methods (Marjanen, 2023; Zichert & Simons, 2026).

Still, several important challenges remain. Embedding shifts can reflect contextual noise rather than substantive semantic drift (Kutuzov et al., 2022), and efforts to classify change types (Cassotti et al., 2024) or assess statistical significance (Liu et al., 2021, pp. 104–113) are still developing. Temporal modeling adds complexity: single models may blur distinctions over time, while separately trained models pose alignment issues. Alternatives like embedding temporal markers or tracking sense clusters remain underused but promising (Periti & Montanelli, 2024). These technical issues also reflect deeper conceptual tensions regarding the operationalization and modeling of conceptual shifts. What counts as a meaningful shift in usage? Current BERT-based approaches typically focus on individual terms and their local context, whereas conceptual history more often concerns the transformation of "semantic fields" (Wevers & Koolen, 2020) or "semantic spaces" (Gavin et al., 2016). This mismatch between words and concepts complicates interpretation and highlights the ongoing need for qualitative judgment.

5.2. Text-level dynamics

Beyond individual tokens, many HPSS questions focus on how larger discursive structures, such as the emergence of ideas, shifts in argumentation, and changes in scholarly communication, develop and circulate over time. Capturing these dynamics requires modeling strategies that span multiple textual scales, from sentences and paragraphs to full arguments, documents, and corpora. We examine LLM-based approaches to these phenomena across three interrelated areas: dynamic topic modeling, scientific novelty detection, and the analysis of citation, influence, and revision. Each addresses a distinct aspect of textual evolution, from the formation of shared vocabularies to the transformation of ideas.

Dynamic topic modeling traces how research themes evolve by grouping texts within temporally segmented corpora based on similarity as estimated in a pretrained embedding space. Recent approaches often rely on text embeddings from the BERT-family, including domain adopted models, such as SciBERT. Among these, BERTopic has become a popular tool for modeling discursive change, combining embedding-based clustering with interpretable topic representations (via c-TF-IDF) and offering built-in "topics over time" workflows. For example, Wang, Chen, et al. (2023) used BERTopic to map interdisciplinary topic trajectories in library and information science, combining topic evolution with diversity and cohesion metrics, while Wang, Downey, and Yang (2023) applied dynamic BERTopic modeling to analyze narratives around AI in international newspapers over 12 years.

Scientific novelty detection aims to identify contributions that diverge meaningfully from prior discourse (Zhao & Zhang, 2025). One approach operationalizes novelty as deviation in embedding space, using text similarity measures to flag outliers. For example, Luo, Lu, et al. (2022) measured distances between text embeddings of research questions and methods, while Just et al. (2024) used cosine similarity across document embeddings. Another approach treats novelty as a learnable feature, using supervised models trained on citation cues or labeled data. Song et al. (2023), for instance, combined BERT embeddings with patent structures to detect emerging technology clusters. Recent work also explores generative methods such as prompt-based scoring (Bornmann et al., 2024; de Winter 2024) or linguistic surprise measures (Vicinanza et al., 2023). Across these approaches, novelty is captured through model-derived signals of difference, whether as embedding-space distance, predicted novelty labels, or surprisal-like scores. Yet for HPSS, this invites caution: epistemic innovation is often incremental, contested, or reframed. LLM-based novelty detection may surface useful candidates, but their significance must be interpreted historically and conceptually, not inferred from distance or scores alone.

Dynamic citation analysis, influence modeling, and revision tracking seek to trace how ideas propagate, shift, and interact across documents through citation, paraphrase, argumentation, or editorial change. Arnaout et al. (2025) used prompt-based learning to classify

"impact-revealing" citations and draft summaries of how a paper's reception appears to evolve across the citing literature. Cheng et al. (2024) linked linguistic similarity to citation lag using article embeddings. Lin et al. (2020) applied BERT-based classification to compare preprints with final publications, showing that more extensive revisions, especially to abstracts and introductions, correlate with eventual acceptance. Gorur et al. (2024) tested generative LLMs on support/attack classification in non-scientific texts, finding strong few-shot performance but challenges with subtle disagreement. Li (2024) used sentence embeddings to trace influence via paraphrased references in historical texts. Jiang et al. (2022) used full-context classifiers to align and compare arXiv versions and categorize revision types.

The methods for tracing text-level dynamics surveyed here remain largely prospective for HPSS, since they have emerged mainly from bibliometrics and NLP. Whether these methods use full-context architectures (including fine-tuned classifiers and embedding-based similarity scoring), generative approaches (few- and zero-shot prompting), or hybrid pipelines, conclusions should always be anchored in inspectable evidence and triangulated through targeted qualitative checks or complementary quantitative measures. Looking ahead, work on LLM-based multi-agent systems suggests a further direction for HPSS, namely agent-based models (ABMs) in which simulated actors interact through generated utterances in dynamic environments (cf. Guo et al., 2024). This echoes Laubichler et al.'s (2019) call to model scientific change via formal representations of agent interaction. While generative ABMs promise more realistic and narrative-rich agent behavior, it remains unclear whether they can yield operationally valid and interpretable accounts of context-sensitive behavior and interaction. Instead, they may shift familiar ABM challenges of calibration, validation, and explanation into more opaque and computationally costly forms (Larooij and Törnberg, 2025).

6. Discussion and lessons

Drawing on the foregoing discussion of models, data, and workflows, we now turn to four lessons to guide the critical and constructive integration of LLMs into HPSS research, and to clarify what responsible use requires in practice.

6.1. Lesson 1: Model selection comes with technical and interpretive trade-offs

Model choice in LLM-based research is not neutral. For HPSS, it requires balancing practical constraints (access, performance, transparency) with interpretive aims and evidential standards. The accessibility–literacy trade-off in Section 2.3 is one central axis of this decision.

Full-context models like BERT variants (e.g., SciBERT) are efficient and well suited to narrowly defined, structured tasks such as classification, retrieval, and embedding-based comparison. They are often smaller, frequently open source, and can often be run locally. Generative models like GPT-5 or Claude-4.5 are more flexible across both structured and open-ended tasks and can compress many steps into a prompt, but are often accessed through proprietary platforms or APIs.

These differences shape what becomes auditable in practice. Embedding-based workflows often leave more intermediate traces (internal parameters, embeddings, similarity scores) and can be easier to inspect end-to-end when run locally. Generative workflows can produce richer narratives, but are harder to reconstruct post hoc and are often more opaque, both in general and through platform constraints.

Hybrid approaches such as RAG and other tool-using setups can combine these strengths by using embeddings to retrieve passages with a documented selection procedure and using generation for tasks whose outputs can be checked against the retrieved evidence. How far this improves auditability depends on whether retrieval results are exposed and prompts and model versions are logged. Multimodal models may further extend this toolkit, though the gain again depends on what evidence is surfaced and whether the pipeline is evaluable and documentable rather than a black box.

6.2. Lesson 2: using LLMs in interpretive research requires LLM literacy

When LLMs enter interpretive workflows, researchers need enough technical and methodological literacy to understand what is being introduced into the evidential chain. Again, Section 2.3's trade-off matters here: ease of use does not reduce the literacy required for responsible interpretation, and pipeline complexity can hide consequential choices even from technically competent users.

LLM outputs include not only fluent text, but also labels, extracted entities, and numerical representations such as CWEs and text embeddings used in downstream analyses. This means literacy has to cover how these outputs are produced, what assumptions shape them, and what errors are plausible. This does not require everyone to become an LLM specialist, but it does require researchers to be able to interrogate the evidential chain. This includes distinguishing full-context from generative models, recognizing how systems are steered through fine-tuning, prompting, or RAG, and seeing how training data, architecture, tool use, and prompts condition results, including when fluent outputs invite over-trust.

LLM literacy is also interpretive. Natural-language outputs require critical reading akin to source criticism and argument analysis (cf. Hemment & Kommers, 2025). Critical engagement therefore spans both technical questions (which system, which tools, which access to sources) and interpretive questions (which categories are imposed, which ambiguities are smoothed away, which alternatives are foreclosed). Without such literacy, there is a risk of epistemic outsourcing. With it, LLMs can extend rather than displace interpretive expertise, provided responsibility for claims remains with HPSS scholars and model contributions remain open to scrutiny.

6.3. Lesson 3: HPSS must define its own datasets, tasks, and evaluation practices

To make LLMs useful for HPSS, scholars need to develop the datasets, tasks, and evaluation practices that guide them. Off-the-shelf benchmarks often assume stable taxonomies, fixed labels, and clear targets, while HPSS materials are historical, genre-diverse, and conceptually contested. Adopting mainstream benchmarks uncritically can therefore import hidden commitments that undermine HPSS perspectives.

Recent critiques underscore these limits. Eriksson et al. (2025) show how benchmarks steer development toward what is easiest to score, with recurring problems such as weak construct validity, leaderboard-driven incentives, and "unknown unknowns" that benchmarks miss. Hemment and Kommers (2025) argue that because LLM outputs function like cultural artifacts, their assessment often requires qualitative judgment, so traditional benchmarking can break down, especially where there is no single ground truth, as with "topic", "discipline", or "conceptual change" in HPSS (cf. Gläser et al., 2017; Meding & Daugs, 2026; Zichert & Simons, 2026).

HPSS therefore needs evaluation regimes that fit interpretive work, and we argue for a plural, explicitly scoped ecosystem of datasets and tasks rather than a one-size-fits-all evaluation setup. This ecosystem should draw on well-documented corpora (time, genre, provenance, editorial history), task formulations that represent ambiguity (for example through multi-labeling, confidence scores, or structured disagreement), and mixed evaluation that combines quantitative baselines with interpretive assessment through expert reading and failure analysis. Evaluation should also articulate clear scope conditions, specifying where a method is expected to work and where it is not, and include robustness checks across varied cases to support bounded generalization. Building this agenda requires sharing norms and collaboration, including partnerships with computer scientists and

alignment with neighboring interpretive fields.

6.4. Lesson 4: LLMs should enhance, not replace, HPSS methodologies

LLMs should be integrated in ways that support, not displace, HPSS's interpretive and reflexive commitments. Even if some reject their use in qualitative research (e.g., Jowsey et al., 2025), LLMs and qualitative methods can interact productively when adopted case by case and paired with shared best practices (cf. Hemment & Kommers, 2025).

How to combine LLMs with HPSS methods depends primarily on two factors: the epistemic role assigned to outputs and how well the model or pipeline has been evaluated for that role. For exploratory uses, such as generating candidate readings, surfacing patterns, or "interpretive sparring", outputs should be treated as proposals rather than conclusions. The aim is to stimulate and guide human reflection, not to outsource judgment. Scaling in this mode is therefore naturally bounded by the pace of human scrutiny and contextualization, and it mainly increases the breadth with which researchers can develop, test, and refine interpretations, rather than simply increasing speed. For justificatory uses, LLM outputs can contribute more directly to evidential support when subtasks are clearly specified, outputs are structured and checkable, and the system has been evaluated and calibrated for the task and domain, ideally with robustness checks and triangulation (Lesson 3). Here scaling can be more straightforward, because a validated procedure can be applied across larger corpora with clearer expectations about error and scope.

In practice, different pipeline components can play different roles. In RAG-based sparring (Section 3.5), generated responses may be exploratory, while retrieval does justificatory work by surfacing passages and delimiting evidence. This suggests a broader principle: evaluate and govern LLM workflows by component and epistemic function, judging exploratory steps by how well they broaden inquiry without misleading it and applying stricter validation and traceability to evidence-bearing steps.

6.5. Beyond methods: infrastructures, values, and responsibilities

Alongside methodological opportunities, HPSS must attend to the infrastructural, ethical, and political conditions under which LLMs are built and adopted. The issue is not only whether outputs are accurate or useful, but how LLM-centered infrastructures reconfigure interpretive workflows and redistribute epistemic authority (cf. Khutsishvili, 2026; Lang, 2026). LLMs sit within a political economy of data extraction, underpaid and hidden human labor, privacy risks, and high environmental costs, driven by weak regulation, platform competition, and current geopolitical and security pressures (Arora et al., 2023; Bender et al., 2021, pp. 610–623; Strubell et al., 2019).

This raises questions of stance and responsibility. While some argue for bans (Jowsey et al., 2025), Guest et al. (2025) propose guiding uptake through research-integrity frameworks such as the Netherlands Code of Conduct, emphasizing honesty, scrupulousness, transparency, independence, and responsibility. Applied to LLMs, this implies disclosure, task-appropriate validation, openness and reproducibility where relevant, safeguards against vendor agendas and conflicts of interest, and sustained attention to legal, social, and environmental harms.

On this view, adopting LLMs for HPSS is not just a methodological choice but an infrastructural and political act that warrants collective governance. Concretely, it calls for clear reporting of tools, datasets, versions, prompts, and preprocessing where relevant, documentation that keeps the evidential chain auditable, and active mitigation of platform dependence, including preferring inspectable components where feasible and defining when constrained use or refusal is required because core integrity conditions cannot be met.

7. Conclusion

LLMs are likely to become part of HPSS research practice, reshaping how scholars move between close reading and corpus-scale analysis. Used well, they can support interpretive analysis while keeping sources available for verification. Used poorly, they can promote "epistemic outsourcing", smoothing ambiguity and historical specificity behind fluent outputs and opaque pipelines. Responsible adoption therefore requires LLM literacy, HPSS-tailored tasks and evaluation, and pipeline-level clarity about what role each component plays. Exploratory outputs should be treated as proposals, while evidence-bearing outputs require validation, robustness checks, and traceability to inspectable materials. These methodological issues also sit within institutional, political, and ethical entanglements that affect research infrastructure and epistemic authority.

We believe that HPSS is well positioned to navigate these tensions. Its long-standing attention to how knowledge and technologies are made, authorized, and contested equips it to scrutinize LLM assumptions, data and labor regimes, and deployment incentives. We therefore advocate a proactive engagement that combines methodological experimentation with critical reflection, extending interpretive capacity without abandoning pluralism or responsibility for justification.

HPSS can also use LLMs as a site for deeper inquiry into explanation, evidence, and responsible use. Historically, it can place them in lineages of quantification and automation that redefined objectivity and expertise. Philosophically, it can clarify the epistemic status of model outputs and connect interpretability debates to earlier work on explanation and epistemic opacity. Sociologically, it can trace the institutions and power struggles shaping LLM development and uptake, from platform dependence to professional authority and public trust. Together, these perspectives position HPSS not only to use LLMs, but to clarify what kind of tools they are and what responsible integration demands.

CRediT authorship contribution statement

Arno Simons: Writing – review & editing, Writing – original draft, Project administration, Data curation, Conceptualization. **Michael Zichert:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Adrian Wüthrich:** Writing – review & editing, Funding acquisition, Conceptualization.

AI statement

During the preparation of the manuscript, we used ChatGPT-4o to assist with improving the language, style, readability, and structure of the text, as well as to help clarify and summarize literature that had already been identified by the authors. All content was subsequently reviewed and edited by the authors, who take full responsibility for the final version of the paper.

Declaration of interest statement

The authors declare no competing interests.

Acknowledgments

We thank Gerd Graβhoff for numerous insightful discussions on the role of large language models in the history, philosophy, and sociology of science, as well as for his helpful comments on an early draft of this paper. We are also grateful to the participants of the interdisciplinary workshop *"Large Language Models for the History, Philosophy, and Sociology of Science"*, held at TU Berlin in April 2025. While the core research presented here predates and extends beyond the event, the workshop provided valuable perspectives that helped sharpen several of the questions addressed in this paper and offered important context for situating our analysis. This work was supported by the European Union

under an ERC Consolidator Grant (Project No. 101044932, "Network Epistemology in Practice (NEPI)"). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Data availability

No data was used for the research described in the article.

References

Ahmadi, E. (2026). In A. Simons, et al. (Eds.), *Exploring disciplinary differences in semantic uniformity: A computational approach to codification* (2026a).

Alampara, N., Schilling-Wilhelmi, M., García, R., et al. (2025). Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science, 5*(10), 952–961.

Arnaout, H., Sternlicht, N., Hope, T., & Gurevych, I. (2025). *In-depth research impact summarization through fine-grained temporal citation analysis*. http://arxiv.org/abs/2505.14838.

Arora, A., Barrett, M., Lee, E., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization, 33*(3), Article 100478.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. http://arxiv.org/abs/1903.10676.

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*.

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. http://arxiv.org/abs/2004.03974.

Bornmann, L., Wu, L., & Ettl, C. (2024). The use of ChatGPT for identifying disruptive papers in science: A first exploration. *Scientometrics, 129*(6), 3593–3598.

Boulanger, C. (2026). The potential of LLMs for constructing a socio-legal knowledge graph. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-4)*. transcript.

Brunila, M., & LaViolette, J. (2022). What company do words keep? Revisiting the distributional semantics of J.R. firth & zellig harris. http://arxiv.org/abs/2205.07750.

Bubeck, S., Chandrasekaran, V., Eldan, R., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. http://arxiv.org/abs/2303.12712.

Buchholz, O., & Grote, T. (2023). Predicting and explaining with machine learning models: Social science as a touchstone. *Studies in History and Philosophy of Science, 102*, 60–69.

Büttner, J. (2026). Why pursue temporally-grounded AI for historical disciplines, and what makes it so challenging? In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-2)*. transcript.

Cassotti, P., Pascale, S. D., & Tahmasebi, N. (2024). Using synchronic definitions and semantic relations to classify semantic change types. http://arxiv.org/abs/2406.03452.

Cassotti, P., & Tahmasebi, N. (2025a). Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics, 13*, 690–708.

Cassotti, P., & Tahmasebi, N. (2025b). A hypothesis-driven framework for detecting lexical semantic change. In C. Bosco, E. Jezek, M. Polignano, & M. Sanguinetti (Eds.), *Proceedings of the eleventh Italian conference on computational linguistics (CLiC-it 2025) (Bd. 4112)*. CEUR https://ceur-ws.org/Vol-4112/#18_main_long.

Chen, Q., Du, J., Allot, A., & Lu, Z. (2022a). LitMC-BERT: Transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19*(5), 2584–2595.

Chen, H., Nguyen, H., & Alghamdi, A. (2022b). Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles. *Scientometrics, 127*(12), 7061–7075.

Cheng, W., Zheng, D., Fu, S., & Cui, J. (2024). Closer in time and higher correlation: Disclosing the relationship between citation similarity and citation interval. *Scientometrics, 129*(7), 4495–4512.

Clark, K., Luong, M.-T., Le, Q., & Manning, C. (2020). Electra: Pre-training text encoders as discriminators rather than generators. http://arxiv.org/abs/2003.10555.

Cohan, A., Feldman, S., Beltagy, I., et al. (2020). *SPECTER: Document-level representation learning using citation-informed transformers*. http://arxiv.org/abs/2004.07180.

Da, N. (2019). The computational case against computational literary studies. *Critical Inquiry, 45*(3), 601–639.

Dagdelen, J., Dunn, A., Lee, S., et al. (2024). Structured information extraction from scientific text with large language models. *Nature Communications, 15*(1), 1418.

Danilova, V. V., Reed, J., Burchell, A., Aangenendt, G., & Söderfeldt, Y. (2026). Zero-shot generation of synthetic historical data with LLMs. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-2)*. transcript.

de Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics, 129*(4), 2469–2487.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. http://arxiv.org/abs/1810.04805.

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify Twitter posts. *Frontiers in Sociology, 7*, Article 886498.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., & Fernandez-Llorca, D. (2025). *Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation*. http://arxiv.org/abs/2502.06559.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. http://arxiv.org/abs/1909.00512.

Falkenberg, M., Galeazzi, A., Torricelli, M., et al. (2022). Growing polarization around climate change on social media. *Nature Climate Change, 12*(12), 1114–1121.

Fergadis, A., Pappas, D., Karamolegkou, A., & Papageorgiou, H. (2021). Argumentation mining in scientific literature for sustainable development. *Proceedings of the 8th workshop on argument mining*.

Gavin, M., Jennings, C., Kersey, L., & Pasanek, B. (2016). Spaces of meaning: Conceptual history, vector semantics, and close reading. In L. F. Klein, & M. Gold (Eds.), *Debates in the digital humanities 2016* (p. 243).

George, L., & Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. *International Journal of Information Technology, 15*(4), 2187–2195.

Gibson, A., Laubichler, M., & Maienschein, J. (2019). Introduction. *Isis, 110*(3), 497–501.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics, 111*(2), 981–998.

Gorur, D., Rago, A., & Toni, F. (2024). Can large language models perform relation-based argument mining?. http://arxiv.org/abs/2402.11243.

Grindrod, J. (2023). Distributional theories of meaning: Experimental philosophy of language. In D. Bordonaba-Plou (Ed.), *Experimental philosophy of language: Perspectives, methods, and prospects* (pp. 75–99).

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. http://arxiv.org/abs/2203.05794.

Gu, Y., Tinn, R., Cheng, H., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH), 3*(1), 1–23.

Guest, O., Suarez, M., Müller, B., et al. (2025). Against the uncritical adoption of "AI" technologies in academia. https://doi.org/10.5281/zenodo.17065099.

Guo, T., Chen, X., Wang, Y., et al. (2024). Large language model based multi-agents: A survey of progress and challenges. http://arxiv.org/abs/2402.01680.

Hangel, N., & ChoGlueck, C. (2023). On the pursuitworthiness of qualitative methods in empirical philosophy of science. *Studies in History and Philosophy of Science, 98*, 29–39.

Harris, Z. (1954). Distributional structure. *Word, 10*(2–3), 146–162.

Hemment, D., & Kommers, C. (2025). *Doing AI differently. Rethinking the foundations of AI via the humanities*. London, UK: The Alan Turing Institute. https://www.turing.ac.uk/news/publications/doing-ai-differently.

Herfeld, C., & Lisciandra, C. (2019). Knowledge transfer and its contexts. *Studies In History and Philosophy of Science Part A, 77*, 1–10.

Hill, M. (2026). The data interview. Reflexive integration of large language models in qualitative content analysis. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-5)*. transcript.

Ho, X., Nguyen, A., Dao, A., et al. (2024). A survey of pre-trained language models for processing scientific text. http://arxiv.org/abs/2401.17824.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language?. In *ACL 2019-57th annual meeting of the association for computational linguistics*.

Ji, Z., Wei, Q., & Xu, H. (2020). Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings, 2020*, 269.

Jiang, C., Xu, W., & Stevens, S. (2022). Edits: Understanding the human revision process in scientific writing. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 9420–9435). Association for Computational Linguistics.

Jowsey, T., Braun, V., Clarke, V., et al. (2025). We reject the use of generative artificial intelligence for reflexive qualitative research (SSRN Scholarly Paper No. 5676462). *Social Science Research Network*. https://doi.org/10.2139/ssrn.5676462

Just, J., Ströhle, T., Füller, J., & Hutter, K. (2024). AI-based novelty detection in crowdsourced idea spaces. *Innovation, 26*(3), 359–386.

Khan, D., Ahmed, I., Ullah, I., & Alwabli, A. (2025). Finding the reference text in citation contexts using attention model. *Service Oriented Computing and Applications, 19*(1), 45–55.

Khutsishvili, K. (2026). AI and the scientist. On the fracture of epistemic authority. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *nderstanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-1)*. transcript.

Kim, K., Kogler, D., & Maliphol, S. (2024). Identifying interdisciplinary emergence in the science of science: Combination of network analysis and BERTopic. *Humanities and Social Sciences Communications, 11*(1), 1–15.

Kleymann, R., Niekler, A., & Burghardt, M. (2022). Conceptual forays: A corpus-based study of "Theory" in digital humanities journals. *Journal of Cultural Analytics, 7*(4).

Kunnath, S., Pride, D., & Knoth, P. (2022). Dynamic context extraction for citation classification. In *The 2nd conference of the Asia-Pacific chapter of the association for*

computational linguistics and the 12th international joint conference on natural language processing, virtual.
Kunnath, S., Pride, D., & Knoth, P. (2023). Prompting strategies for citation classification. *Proceedings of the 32nd ACM international conference on information and knowledge management*.
Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. http://arxiv.org/abs/1806.03537.
Kutuzov, A., Velldal, E., & Øvrelid, L. (2022). Contextualized embeddings for semantic change detection: Lessons learned. *Northern European Journal of Language Technology, 8*(1). Article 1.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.
Lang, S. (2026). Critical concerns for using LLMs in the (computational) humanities and beyond. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-1)*. transcript.
Larooij, M., & Törnberg, P. (2025). Do large language models solve the problems of agent-based modeling? A critical review of generative social simulations. http://arxiv.org/abs/2504.03274.
Laubichler, M., Maienschein, J., & Renn, J. (2019). Computational history of knowledge: Challenges and opportunities. *Isis, 110*(3), 502–512.
Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.
Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.
Leydesdorff, L., Ràfols, I., & Milojević, S. (2020). Bridging the divide between qualitative and quantitative science studies. *Quantitative Science Studies, 1*(3), 918–926.
Li, L. (2024). Tracing the genealogies of ideas with sentence embeddings. In M. Hämäläinen, E. Öhman, S. Miyagawa, et al. (Eds.), *Proceedings of the 4th international conference on natural language processing for digital humanities* (pp. 9–16). Association for Computational Linguistics.
Liesegang, L., & Gläser, J. (2026). Supporting citation context analysis with large language models raises questions that should have been asked 40 years ago. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-6)*. transcript.
Lin, J., Yu, Y., Zhou, Y., et al. (2020). How many preprints have actually been printed and why: A case study of computer science preprints on arXiv. *Scientometrics, 124*(1), 555–574.
Liu, Y., Medlar, A., & Glowacka, D. (2021). Statistically significant detection of semantic shifts using contextual word embeddings. *Proceedings of the 2nd workshop on evaluation and comparison of NLP systems*.
Liu, Y., Ott, M., Goyal, N., Du, J., et al. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. http://arxiv.org/abs/1907.11692.
Liu, P., Yuan, W., Fu, J., Jiang, Z., et al. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 195:1–195:35.
Lucy, L., Dodge, J., Bamman, D., & Keith, K. (2023). Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 6929–6947). Association for Computational Linguistics.
Luo, Z., Lu, W., He, J., & Wang, Y. (2022a). Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics, 16*(2), Article 101282.
Luo, R., Sun, L., Xia, Y., et al. (2022b). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics, 23*(6). bbac409.
Ma, B., Zhang, C., Wang, Y., & Deng, S. (2022). Enhancing identification of structure function of academic articles using contextual information. *Scientometrics, 127*(2), 885–925.
Malaterre, C., & Lareau, F. (2026). Epistemic framings in science. Charting scientific knowledge with embeddings and LLMs. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-3)*. transcript.
Meding, H., & Daugs, A. (2026). On the use and limitations of large language models in historical scholarship. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-1)*. transcript.
Marjanen, J. (2023). Quantitative conceptual history: On agency, reception, and interpretation. *Contributions to the History of Concepts, 18*(1), 46–67.
Mikolov, T., Sutskever, I., & Chen, K. (2024). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.
Nicholson, J., Mordaunt, M., Lopez, P., et al. (2021). Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies, 2*(3), 882–898.
Periti, F., Dubossarsky, H., & Tahmasebi, N. (2024). (Chat)GPT v BERT: Dawn of justice for semantic change detection. http://arxiv.org/abs/2401.14040.
Periti, F., & Montanelli, S. (2024). Lexical semantic change through large language models: A survey. *ACM Computing Surveys, 56*(11), 282:1–282:38.
Pham, C., Hoyle, A., Sun, S., & Iyyer, M. (2023). TopicGPT: A prompt-based topic modeling framework. http://arxiv.org/abs/2311.01449.
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.
Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. http://arxiv.org/abs/1908.10084.
Roman, M., Shahid, A., Khan, S., et al. (2021). Citation intent classification using word embedding. *IEEE Access, 9*, 9982–9995.
Rosin, G., & Radinsky, K. (2022). *Temporal attention for language models*. http://arxiv.org/abs/2202.02093.
Scharnhorst, A., Yang, H., Touber, J., Ferguson, K., Mayr, P., & Tykhonov, V. (2026). Co-creation of AI technology, empowering curators of cultural heritage information and guarding research commons. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-5)*. transcript.
Schlattmann, R., Kaye, A., & Vogl, M. (2026). From source to structure. Extracting knowledge graphs with LLMs. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-4)*. transcript.
Shao, W., Ji, P., Fan, D., et al. (2024). *Astronomical knowledge entity extraction in astrophysics journal articles via large language models*. http://arxiv.org/abs/2310.17892.
Simons, A. (2024a). *Astro-HEP-BERT: A bidirectional language model for studying the meanings of concepts in astrophysics and high energy physics*. http://arxiv.org/abs/2411.14877.
Simons, A. (2024b). *Meaning at the planck scale? Contextualized word embeddings for doing history, philosophy, and sociology of science*. https://arxiv.org/abs/2411.14073.
Simons, A., Arnaout, H., & Gurevych, I. (2026). Reconstructive citation context analysis using large language models. A roadmap. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-6)*. transcript.
Simons, A., Wüthrich, A., Zichert, M., & Graßhoff, G. (2026). *Understanding science with large language models? Potentials for the history, philosophy, and sociology of science*. transcript.
Song, B., Luan, C., & Liang, D. (2023). Identification of emerging technology topics (ETTs) using BERT-based model and sematic analysis: A perspective of multiple-field characteristics of patented inventions (MFCOPIs). *Scientometrics, 128*(11), 5883–5904.
Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3645–3650).
Taylor, R., Kardas, M., Cucurull, G., et al. (2022). Galactica: A large language model for science. http://arxiv.org/abs/2211.09085.
Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.
Tykhonov, V., Yang, H., Mayr, P., et al. (2025). Chatting with papers: A hybrid approach using LLMs and knowledge graphs. http://arxiv.org/abs/2505.11633.
Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.
Vicinanza, P., Goldberg, A., & Srivastava, S. (2023). A deep-learning model of prescient ideas demonstrates that they emerge from the periphery. *PNAS Nexus, 2*(1), 275.
Wang, Z., Chen, J., Chen, J., & Chen, H. (2023b). Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, 1–26.
Wang, W., Downey, J., & Yang, F. (2023a). *AI anxiety? Comparing the sociotechnical imaginaries of artificial intelligence in UK, Chinese and Indian newspapers*. Global Media and China.
Wang, H., Prakash, N., Hoang, N., et al. (2023). Prompting large language models for topic modeling. In *2023 IEEE international conference on big data (BigData)* (pp. 1236–1241).
Wei, J., Tay, Y., Bommasani, R., et al. (2022). *Emergent abilities of large language models*. http://arxiv.org/abs/2206.07682.
Wevers, M., & Koolen, M. (2020). Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53*(4), 226–243.
Wu, J., Gan, W., Chen, Z., et al. (2023). Multimodal large language models: A survey. In *2023 IEEE international conference on big data (BigData)* (pp. 2247–2256).
Wu, X., Nguyen, T., Zhang, D., et al. (2024). FASTopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. http://arxiv.org/abs/2405.17978.
Zhang, Q., Ding, K., Lyv, T., et al. (2024). Scientific large language models: A survey on biological & chemical domains. http://arxiv.org/abs/2401.14656.
Zhang, Z., Fang, M., Chen, L., & Namazi-Rad, M.-R. (2022). Is neural topic modelling better than clustering? An empirical study on clustering with contextual embeddings for topics. http://arxiv.org/abs/2204.09874.
Zhao, Y., & Zhang, C. (2025). A review on the novelty measurements of academic papers. *Scientometrics, 130*(2), 727–753.
Zhu, Y., Yuan, H., Wang, S., et al. (2024). Large language models for information retrieval: A survey. http://arxiv.org/abs/2308.07107.
Zichert, M., Simons, A., & Wüthrich, A. (2025). Expanding conceptual histories: Using contextualized word embeddings for the history and philosophy of the virtual particle concept. *Computational Humanities Research, 1*. Article e16.
Zichert, M., & Simons, A. (2026). From early digital methods to LLMs. Computational conceptual history of scientific concepts. In A. Simons, A. Wüthrich, M. Zichert, & G. Graßhoff (Eds.), *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science (part-3)*. transcript.